Aryan Khatri

Jaipur, Rajasthan, India
linkedin.com/in/aryan-khatri github.com/EHyashu Portfolio
khatriaryan880@gmail.com — +91-8595165654

Professional Summary

GenAI & ML Engineer specializing in production-grade RAG systems, LLM applications, and scalable AI backends. Experienced in designing agentic architectures, evaluation pipelines, and real-time systems using FastAPI, Docker, and vector databases. Strong focus on retrieval quality, latency optimization, and deployable AI systems.

Technical Skills

Machine Learning: Regression, Classification, Clustering, Feature Engineering, Model Evaluation
Languages: Python, SQL
Generative AI / LLMs: FastAPI, LangChain, LlamaIndex, Streamlit, Agno, Prompt Engineering, RAG, Transformers, Evaluation (RAGAS)
Tools: Docker, MLflow, Git, AWS
Databases: Qdrant, ChormaDB, PostgreSQL, Redis
Libraries: Scikit-learn, Pandas, NumPy, OpenCV, Pytorch, spaCy, Matplotlib

Experience

Machine Learning Intern — Intrainz Mar 2025 – May 2025

- Built a fraud detection system to identify fake screenshots using image processing and ML
- Extracted features (metadata, pixel inconsistencies, compression artifacts) to detect tampering
- Automated preprocessing and evaluation pipelines, reducing effort by 40% using MLflow

Projects

High-Fidelity Domain-Specific RAG Pipeline & Evaluation Framework GitHub

- Architected end-to-end RAG pipeline using LangChain + Groq (LLaMA) for technical document QA
- Implemented advanced chunking (1000 tokens, overlap 200) using RecursiveCharacterTextSplitter
- Integrated ChromaDB vector store with Jina embeddings for dense semantic retrieval
- Built RAGAS evaluation pipeline measuring Faithfulness, Context Precision, Answer Relevancy
- Engineered robust API layer with retry logic and structured JSON outputs for stability

UniGuide — Agentic RAG System GitHub

- Designed multi-agent architecture (Router, Retrieval, Web, Memory, Answering agents)
- Implemented hybrid retrieval (Qdrant + keyword filtering) with HyDE-based query expansion
- Built real-time chat backend (FastAPI + WebSockets) with Redis caching for low latency
- Developed evaluation pipeline using RAGAS and LLM-generated gold datasets
- Optimized inference using small/large LLM routing (Groq LLaMA) for cost-performance tradeoff

AI Interviewer — Multi-Agent Evaluation System GitHub

- Built AI interview platform using multi-agent LLM orchestration for HR, technical, and coding rounds
- Implemented dynamic interview flow control, scoring, and structured feedback generation
- Integrated ML-based scoring and WebSocket backend for real-time interaction
- Added behavioral analysis using computer vision for engagement insights

WebCam — Real-Time Motion Detection System GitHub

- Developed real-time video analytics pipeline using OpenCV (frame differencing, contour tracking)
- Built live monitoring dashboard with alert system using Streamlit
- Improved detection robustness using adaptive thresholding and noise filtering

Education

B.Tech in Artificial Intelligence & Data Science
Poornima University 2023 – 2027

Achievements

Finalist — LNMIIT Hackathon | Winner — IIT Delhi Modeling Competition